

Semantic Simplification and Regional Translation of Legal Texts using LLMs and Sentence Segmentation

Shivappa M Metagar¹, Vinayak V Pottigar² and Farooque R Sayyed³

^{1,3}Assistant Professor, CSE Department Walchand Institute of Technology, Solapur, Maharashtra, India-413006
Email: shivametagar@gmail.com, sayyed.farooque999@gmail.com

²Assistant Professor, IT Department Walchand Institute of Technology, Solapur, Maharashtra, India-413006
Email: vinayakpottigar2003@gmail.com

Abstract— Legal documents related to the registration or violation of intellectual property rights (IPR) often feature complex sentence structures, ambiguous expressions, and a widespread use of legal jargon, making the content difficult to understand for non-experts. This case is in particular obtrusive in a country like India, where numerous languages are spoken and written, within a context of linguistic and cultural range, and where each character has awesome rights. In particular, within the case of shielding such rights, the availability of clean criminal records is crucial. In this article, we present an optimized version primarily based at the T5-base structure, a neural network designed to simplify texts within the context of highbrow assets regulation in India. The model rewrites the unique prison text right into a simplified shape even as maintaining the criminal meaning of the language. It has been skilled with the Indian prison report Corpus (ILDC), which includes over 35,000 court instances and their authentic choices as training examples. Moreover, it converts criminal textual content into various neighborhood languages to decorate accessibility for every Indian citizen. The version's assessment, at the same time as simplifying complex felony phrases and reverting them to extra problematic bureaucracy, has shown a huge reduction in linguistic complexity at the same time as preserving the criminal informational content material intact, and the model generates the specific report in real-time, translating it into distinctive languages.

Index Terms— Legal Text, Indian legal files Corpus (ILDC), Large Language Model (LLM).

I. INTRODUCTION

In cutting-edge unexpectedly evolving and interconnected international of generation and innovation, intellectual property rights (IPR) function the cornerstone for shielding invention, entrepreneurship, and organizational improvement. As global industries boost and convey an increasing extent of intellectual belongings within the shape of patents, copyrights, and trademarks, the call for protection will intensify. In a various united states like India, wherein over 22 authentic languages are spoken along numerous dialects, getting access to and information complicated criminal documents poses a large assignment for many. The technical language utilized in IPR-associated files regularly hinders comprehension for individuals from various instructional and cultural backgrounds. Consequently, there's a urgent want for simplification and subsequent translation into nearby languages.

The proposed initiative ambitions to bridge this essential hole by growing an advanced intelligent platform designed to simplify prison files specifically related to intellectual belongings rights (IPR) law and facilitate IPR promotion and management (CIPAM), even as also translating them into local Indian languages. this may be carried out through natural Language Processing (NLP) technology, utilizing a cultured large Language model (LLM) that employs a paraphrase based on a T5-base version to perform those obligations. The language utilized in documents related to highbrow assets rights (IPR) frequently poses a task for individuals from different educational and cultural backgrounds. Therefore, there may be a pressing need for simplification and subsequent translation into nearby languages. The proposed initiative goals to bridge this critical hole by means of creating a revolutionary realistic platform designed to simplify felony files specifically relating IPR law and facilitate the merchandising and control of IPR (CIPAM), at the same time as also ensuring seamless translation into regional Indian languages.

II. LITERATURE SURVEY

The significance of successful communication throughout numerous cultures is more applicable now than ever in our interconnected and globalized globally (see authors Yasir Abdelgadir Mohamed et al. [1]). Language is an important mode through which people, businesses, and countries are connected, therefore necessitating well executed and accurate translation systems. The motive of this evaluation paper is to contribute to the commune on AI-driven language translation. This review of the literature will seriously interact with the literature that already exists, literature that advocates for contrary perspectives, in addition to perceive possibilities for innovation, highlight regions of development, and perceive limitations.

The authors, Shounak Paul et al. [2] present that legal systems throughout the sector are being inundated with cases and documents at historic charges.

Therefore, there is increasing urgency to develop Natural Language Processing (NLP) and Machine Learning (ML) methods for automating the processing and understanding of legal documents to help lighten the load on the legal system. But evaluating and comparing diverse NLP models in the legal domain presents significant complications. The current paper tackles this challenge by proposing the Indian Legal Text Understanding and Reasoning (IL-TUR) benchmark, which offers monolingual (English, Hindi) and multilingual (9 Indian languages) domain-specific tasks focusing on multiple aspects of the Indian legal system, with an emphasis on understanding and reasoning with Indian legal documents.

Rahul C. Kore and colleagues at Dell Technologies [3] emphasized that reviewing legal documents could be tiresome at times as it could require domain knowledge dealing with legal proof. It's difficult to read the entire document without losing or misplacing contradicting or additional important sentences. Studies show on average people review meaningfully large amounts of present legal documents than a person could read. With the increasing number of legal documents, it would be more efficient to get the key-indicating information from the document to avoid reading the document in its entirety. The authors explored achieving this by trying to understand the vast legal document in a determinate time period. Summarization promotes a level of flexibility and convenience for the reader.

The authors Li et al. [4] emphasized that intralingual translation is a significant linguistic mediation that facilitates meaning construction, with paraphrasing being a prominent type of intralingual translation. While also employed in a variety of contexts, such as language learning and translation training with textual and sentential properties that are distinct from the original text. Documenting cognitive processes of paraphrasing can be compared to interlingual translation, and, therefore, we carried out a comparative study to examine cognitive processes at both the text and sentence levels while engaging in translation and paraphrasing utilizing eye-tracking and retrospective protocols.

Anggawijaya et. al. [5] State that When translating technical language from English to Indonesian, the paraphrasing technique is adopted as a way to translate a technical term so that the meaning will be accessible to the reader. In this article, we will demonstrate the use of paraphrasing in the context of a systematic article entitled "How individuals exchange Language," mainly focused on the majority. this newsletter is primarily based on a qualitative approach through analysis of a way inside the translation of the source textual content, with two versions of translation being considered as steps: a default translation and another translation suitable for the public. Each excerpt has a discussion of paraphrasing used in simplifying technical terms.

The author Jianwei Zheng et al. [6] indicate that terminology is an important representation of the professionalism of academic discourses. Faithful reconstruing of the source text's terminology keeps focal in translation studies of terminology, whereas terminology shift and flexible translation of the supply textual content's terminologies have received little interest. primarily based at the processing manuscripts of three

books from the ebook collection accumulated Translations of Belt and avenue research, this paper describes the translators' use and some feasible outcomes of terminology shift.

The authors, Surbhu Sharma et al. [7], articulate that many translation systems and applications, including textual translation and speech systems, use it effectively, although somewhat limited in terms of grammatical correctness and completeness. Device translation has been around for a long term and acts as a bridge for move-language conversation in these day's wired international. Statistical machine Translation (SMT) makes use of methods to translate these statements right into a goal language without changing the means.

As mentioned by the authors Dechao Li et al. [8], translation and paraphrasing abound in second language (L2) communication in general, and they are effective techniques in second language acquisition (SLA). Although there are a good number of studies comparing their similarities and differences from a process-oriented perspective the relationship between product great has been much less examined probably because it is not clean to evaluate the pleasant of translation or paraphrasing or an affordable technique to doing so has been too subjective, unidimensional.

The author Guifang Wang alongside along with his colleagues [9] explained abstract: The purpose of this paper is to study the interpretation paths of chinese language criminal texts in the difficulty of translation studies, deliberating cultural, linguistic, and felony contexts. The item opens with a very deep description of the chinese language language language felony textual content, emphasizing a number of the characteristics of the textual content. The dialogue then progresses to explain the theoretical foundation of translation studies, focusing on equivalence and the specific strategies of translation.

The usage of system translation has elicited giant research in language pairs; lots paintings has been finished on its effectiveness in producing any and all translations, yet it's miles obtrusive that the translation of specialized languages has merited limited have a look at. The current have a look at examines the effectiveness of using device translation while translating prison texts and assesses the exceptional of device translated felony texts from Arabic into English.

III. PROPOSED WORK

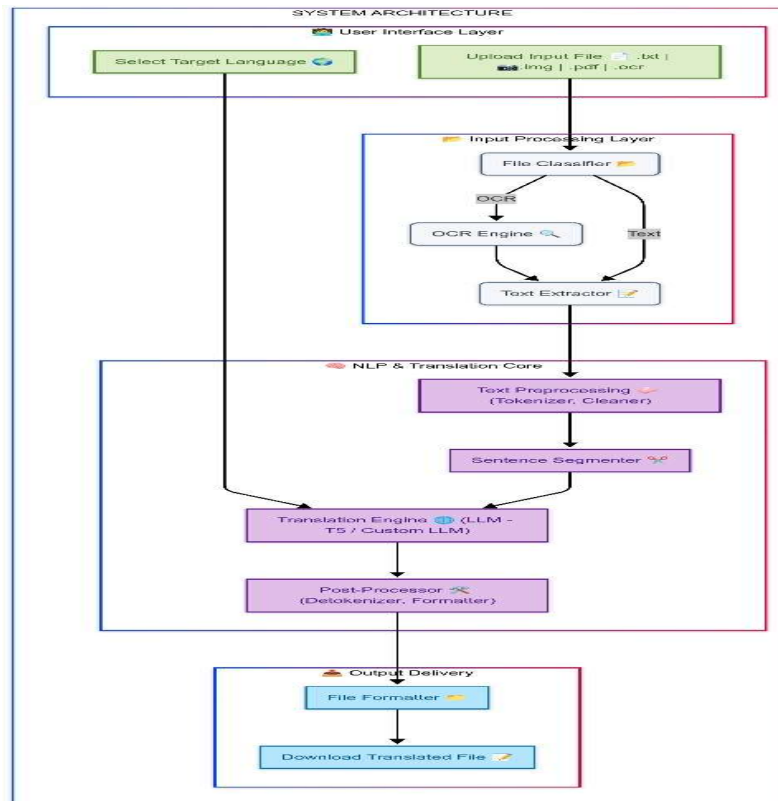


Fig 1; System Architecture

The language used in documents related to intellectual property rights (IPR) often poses a challenge for individuals from different educational and cultural backgrounds. Consequently, there is a pressing need for simplification and subsequent translation into local languages. The proposed initiative aims to bridge this critical gap by creating an innovative practical platform designed to simplify legal documents specifically pertaining to IPR law and facilitate the promotion and management of IPR (CIPAM), while also ensuring seamless translation into regional Indian languages.

A key aspect of this platform is its accessibility; you no longer have to worry about the type of premium membership or software download; access to this service is always just a click away and it is free! This tool ensures that translations reflect the intent and meaning of the source material when performing a legal translation; it ensures that warranty translations maintain legal clauses.

IV. METHODOLOGY AND ARCHITECTURE

This system, evolved to simplify and translate legal files related to CIPAM and IPR, often makes use of natural language processing and device learning technology for real-time document translation. The structure of the system facilitates the processing of numerous varieties of input files through its more than one modules, enabling correct translation into distinctive nearby languages.

The technique outlined within the paper is structured into several wonderful degrees, as illustrated within the accompanying flowchart

- File Type Selection: The process starts with the identification of the input document type. users have the choice to put up numerous codec's, consisting of textual content files (.txt), scanned documents (utilizing OCR), or PDFs. this flexibility allows the platform to deal with a huge range of file sorts, permitting users to paintings with any report they own.
- File Upload and Processing: Once the user has selected the file kind, they continue to add the report or textual content meant for translation. This initiates the pre-processing segment, at some stage in which the file or textual content is wiped clean and organized for similarly processing. For scanned documents, the gadget employs Optical individual reputation (OCR) generation to convert the uploaded picture into device-readable textual content that the machine can process.
- Language Selection: Following the pre-processing of the document, users can choose their preferred local language for translation. The system offers a variety of options to meet the needs of its target user base. Currently, the system is capable of translating into several Indian languages, such as Hindi, Marathi, and Telugu.
- Paraphrasing and Translation: The primary objective of the system is to simplify the criminal content in the file thru a textual content-to-textual content transformation, basically paraphrasing the textual content the use of a finely-tuned language version. This paraphrasing process utilizes a large Language model (LLM), in particular a ChatGPT paraphraser based on T5-base, which has been first-class-tuned on a dataset of legal files (Indian prison documents Corpus).
- Creation and Download of Translations: as soon as the paraphrasing and translation tactics are finalized, the device is able to producing a file that the person can subsequently download. The consumer has the choice to down load the finalized content material in their desired document format, that means that the translated record is prepared for use upon download, each legally and for my part.
- Summary: After the user has downloaded the translated transcript file, they have completed their interaction with the system, marking the conclusion of this workflow.

Algorithmic Steps:

To streamline the architecture into a systematic algorithm, follow these steps:

Step 1: Initiate the platform by selecting the desired file type.

Step 2: Choose from the available file formats: txt, ocr, or .pdf.

Step 3: Upload the file intended for translation onto the platform.

Step 4: Preprocess the document by cleaning the text and converting any scanned text from the OCR file into a machine-readable format.

Step 5: Select the appropriate regional language for translation, such as Hindi, Marathi, Telugu, etc.

Step 6: Utilize the LLM to paraphrase and simplify legal clauses and terminology.

Step 7: Translate the paraphrased content and legal clauses into the chosen local regional language.

Step 8: Generate the required output.

Step 9: Enable the user to download the document in a readable format in the selected regional language.

Step 10: Conclude the process.

V. RESULTS AND COMPARISON

In order to evaluate the effectiveness of our project in enhancing translation performance, we performed a straightforward comparison analysis with Google Translate, utilizing identical input texts. We translated these texts into the local languages of Hindi, Marathi, and Kannada, with a particular focus on examining the translation of legal nuances.

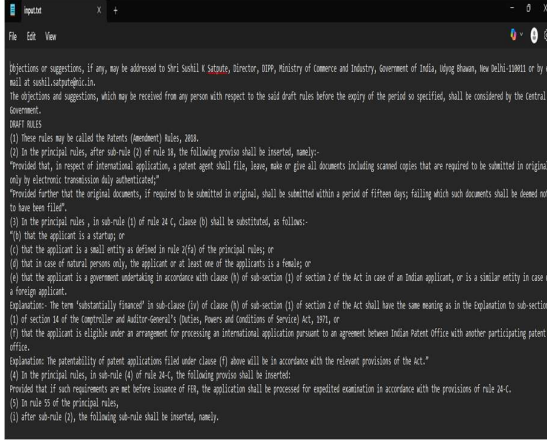


Fig 2: Original Text

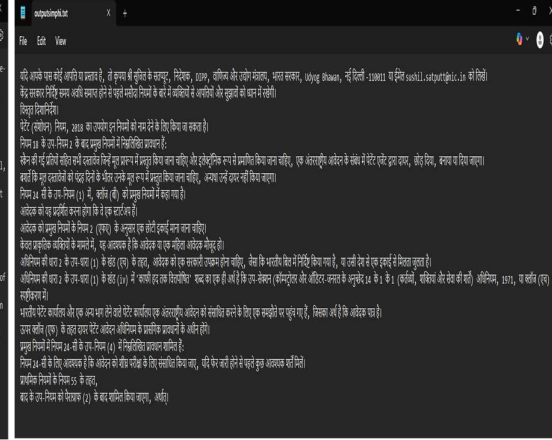


Fig 3: Simplified TextIn Hindi

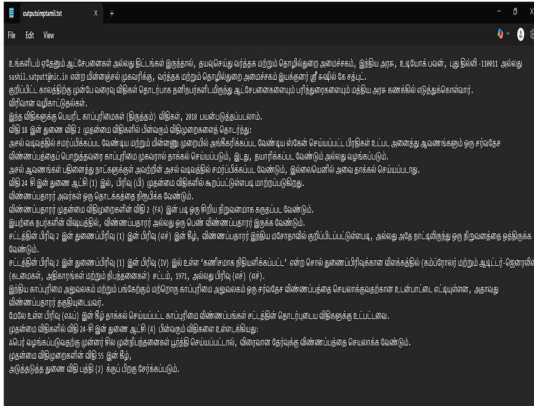


Fig 4: Simplified Text In Tamil

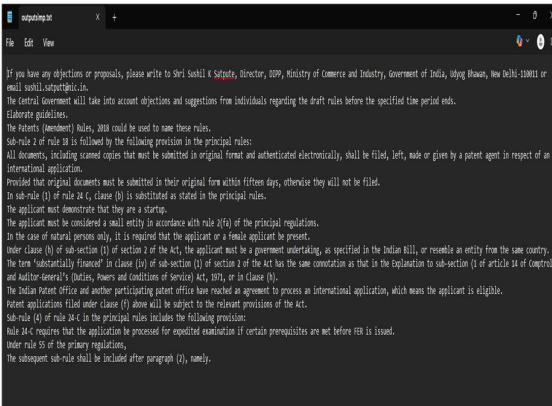


Fig 5: Simplified Text In English

VI. CONCLUSIONS

The effort to translate and simplify complex legal documents, particularly those related to IPR and CIPAM, represents not just a technological milestone but also a significant social initiative. This endeavour promises to greatly enhance the accessibility of legal knowledge across diverse linguistic communities in our nation. Soon, individuals from various backgrounds will have the opportunity to engage with the legal frameworks that impact their lives, as well as the intricate laws that influence our society and everyday experiences. With the advancements in Natural Language Processing and modern machine learning techniques, our team at the CLS has developed an intelligent system capable of delivering these translations in real time, ensuring that the essence of legal intent is accurately conveyed to those without specialized training. Through meticulous pre-processing, the system generates refined outputs that are ready for practical application.

REFERENCES

- [1] Yasir Abdelgadir Mohamed, (Member, IEEE), Akbar Khanan, Mohamed Bashir et al, "The Impact of Artificial Intelligence on Language Translation: A Review". Received 24 January 2024, accepted 4 February 2024, Digital Object Identifier 10.1109/ACCESS.2024.3366802.

- [2] Shounak Paul, Abhinav Joshi, Akshat Sharma, Pawan Goyal, Saptarshi Ghosh, Ashutosh Modi ajoshi, akshatsh, ashutoshm@cse.iitk.ac.in, "IL-TUR: Benchmark for Indian Legal Text Understanding and Reasoning". Received 7 Jul 2024. arXiv:2407.05399v1 [cs.CL].
- [3] Rahul C Kore, Prachi Ray, Priyanka Lade, Amit Nerurkar (Vidyalankar Institute of Technology, Mumbai University, Mumbai, India) , "Legal Document Summarization Using NLP and ML Techniques". Volume 9 Issue 05 May 2020, Page No. 25039-25046 ISSN: 2319-7242 DOI: 10.18535/ijecs/v9i05.4488.
- [4] DECHAO LI. Correspondence Author. "Global and Local Cognitive Processing in Paraphrasing and Translation –A Comparative Study using Eye-tracking and Retrospective Protocol". on 17 July 2023, [https://doi.org/10.1016/S0166-4115\(08\)61493-4](https://doi.org/10.1016/S0166-4115(08)61493-4).
- [5] Muhammad Hendy Anggawijaya, Dimas Adika, "Enhancing Target Text Comprehension for Lay Audience through Paraphrasing". Received e-ISSN: 2798-950X | ISSN: 2798-9518 Vol. 3, No. 1, June 2023, pp.1-14, <https://doi.org/10.33830/humaya.v3i1.4282>.
- [6] Jianwei Zheng, Fuzhou University, "Terminological Shift in Academic Translation: A Descriptive Study Based on Processing Manuscripts". on 01 August 2024. DOI: 10.12339 / j. Issn. 1673-8578. 2023. 02. 007.
- [7] Shurbi Sharma, "A Transformer based approach using LSTM and Paraphrase reference to Translate English Text into Hindi". Publication Date: August 14th, 2023, DOI: <https://doi.org/10.21203/rs.3.rs-3230686/v1>.
- [8] Dechao Li, Tianyi Han, Xingcheng Ma And Nan Hu. "Comparing product quality between translation and paraphrasing: Using NLP-assisted evaluation frameworks". Published: 25 November 2022, DOI 10.3389/fpsyg.2022.1048132.
- [9] Pengbin Zhang, Guifang Wang, "Research on Translation Path of Chinese Legal Texts". Received ISSN: 2770-6702 | Vol. 12, No. 2, 2024.
- [10] Eman Rashed Alkathery, "Google Translate Errors in Legal Texts: Machine Translation Quality Assessment". Received: 11/07/2022 Accepted: 1/31/2023 Published: 02/24/2023, AWEJ for Translation & Literary Studies, Volume 7, Number 1. February 2023 DOI: <http://dx.doi.org/10.24093/awejtls/vol7no1.16>.